Statistics
in Medicine

Research Article

Received 10 July 2012, Accepted 26 May 2013 Published online 3 July 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5888

Integrative variable selection via Bayesian model uncertainty

M. A. Quintana[a] and D. V. Conti[b*†]

We are interested in developing integrative approaches for variable selection problems that incorporate external knowledge on a set of predictors of interest. In particular, we have developed an integrative Bayesian model uncertainty (iBMU) method, which formally incorporates multiple sources of data via a second-stage probit model on the probability that any predictor is associated with the outcome of interest. Using simulations, we demonstrate that iBMU leads to an increase in power to detect true marginal associations over more commonly used variable selection techniques, such as least absolute shrinkage and selection operator and elastic net. In addition, iBMU leads to a more efficient model search algorithm over the basic BMU method even when the predictor-level covariates are only modestly informative. The increase in power and efficiency of our method becomes more substantial as the predictor-level covariates become more informative. Finally, we demonstrate the power and flexibility of iBMU for integrating both gene structure and functional biomarker information into a candidate gene study investigating over 50 genes in the brain reward system and their role with smoking cessation from the Pharmacogenetics of Nicotine Addiction and Treatment Consortium. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: Bayes factors; informative model space prior; genetic association studies; group variable selection

1. Introduction

Throughout this paper, we are focused on developing integrative variable selection techniques for high-dimensional problems. These problems arise in diverse areas from genetic and environmental epidemiology to predicting the stock market, wine prices, and scouting professional athletes. In many of these areas, recent technological advances allow for the collection of massive datasets. Traditional analyses relied heavily on an expert to meticulously sift through the data deliberately building models for inference based on contextual knowledge. However, with the shear amount of factors available for evaluation, it is often impracticable to build a model in this way. This analysis bottleneck has lead to a shift to data-driven and computer-based data mining approaches often at the expense of the context in which the statistical model may be relevant. The methods developed herein aim to integrate these approaches to allow for high-dimensional model search while incorporating multiple sources of data.

As an example, we are motivated by the application of these integrative techniques to genetic association studies. It is now feasible to obtain genotype information on millions of variants across the genome, and these studies are seeing a dramatic increase in popularity for numerous complex diseases. Given the vast amount of potential predictors of interest, most analyses treat each variant as independent and limit the investigation to only marginal associations. The simplistic analyses of these studies lead to power limitations due to small marginal effect sizes for complex diseases and the strict thresholds imposed to determine marginal associations. Also, it is becoming increasingly likely that combinations of variants, including rare variant loads or interactions between variants, may be important. Thus, there is a growing interest in determining a set of genetic markers associated with an outcome of interest. Although data-driven methods are available for selecting sets of markers, they ignore any prior information or fail to integrate additional data sources that may refine the set selection. Specifically, because of the power and

[a]*Berry Consultants, Austin, TX, U.S.A.*
[b]*Department of Preventive Medicine, Division of Biostatistics, University of Southern California, Los Angeles, CA, U.S.A.*
**Correspondence to: D. V. Conti, Department of Preventive Medicine, Division of Biostatistics, University of Southern California, Los Angeles, CA, U.S.A.*
[†]*E-mail: dconti@usc.edu*

computational limitations of genetic association studies and the increasing availability of complementary data sources, it is becoming even more necessary to integrate multiple sources of data to discover multivariate genetic profiles for complex disease.

Given the inability to quantify external information formally into prior beliefs for a large number of predictor variables of interest, external biological knowledge is often ignored completely in agnostic scans (e.g., genome-wide association study). If the knowledge is used, it is mostly limited to the design phase of a study in which a set of candidate genes and/or variants of interest are specifically chosen to be genotyped. With this in mind, we focus on using a set of factors to help guide the selection of variables empirically. These factors can be thought of as a set of predictor-level covariates that reflect higher-level relationships between the predictors. To construct predictor-level covariates in genetic association studies, one may use previously existing and curated ontological data, such as of the Gene Ontology [1], or other various pathway ontology databases such as KEGG [2], BioCyc [3], Reactome [4], and PANTHER Pathways [5]. Another example may be to construct the predictor-level covariates from high-density metabolomic, transcriptomic, and proteomic data collected on individuals. In each case, collections of variables are defined as conditionally exchangeable groups based on the covariate specification and the estimated relative importance of that characterization on the outcome of interest.

Previous methods incorporating predictor-level covariates have focused on informing estimation [6–13]. Unlike these methods, we focus on model selection and hypothesis testing incorporating external information into the probability that any predictor variable is associated with the outcome of interest. Other penalized regression methods such as least absolute shrinkage and selection operator (lasso) [14] and elastic net [15] are common variable selection techniques. Although elastic net does not explicitly incorporate predictor-level external information, it does allow for the correlation structure of the predictors to guide the variable selection procedure by introducing a hybrid between the L_1 penalty of lasso and the L_2 penalty of ridge regression [16]. More recent methods such as the group lasso [17] and group bridge [18] explicitly allow for the inclusion of group-based predictor covariates (essentially dichotomous covariates) to inform the selection. The main critique of group lasso is that it selects all or none of the predictor variables within a group by introducing a group-level penalty. The group bridge overcomes this drawback by also introducing a variable-level penalty to allow for group-level and within-group variable selection.

The Bayesian model uncertainty (BMU) framework has been shown to be extremely powerful within variable selection problems. In particular, for genetic association studies, Wilson *et al.* [19] have demonstrated that the BMU approach leads to an increase in power to detect truly associated variants over more commonly used variable selection techniques such as lasso, stepwise regression, and marginal multiplicity adjusted approaches. On the basis of the power and flexibility of the BMU framework for variable selection problems, we are interested in extending this framework to incorporate external predictor-level knowledge. Within the BMU framework, Chipman [20] and Conti *et al.* [21, 22] described an informative prior for related predictors that introduces dependencies between higher-order interaction terms and their 'parent' terms. Similarly, Stingo *et al.*, Baurley *et al.*, and Li and Zhang [23–25] described methods to incorporate a known graphical structure into the prior probability that each variable is associated. We wish to build upon these priors and develop a more general approach for incorporating external information into the BMU framework by introducing iBMU. Within iBMU, we introduce a second-stage hierarchical probit model on the probability that each predictor variable is associated with the outcome of interest that is a function of a set of predictor-level covariates and their empirically estimated effects. Unlike the group penalized regression approaches of Yuan and Lin [17] and Huang *et al.* [18] that account only for the inclusion of dichotomous covariates, our approach has been created within a general framework that allows for the integration of multiple sources of external information within the form of both continuous and dichotomous covariates.

The rest of the paper is organized as follows. Section 2 gives an overview of the BMU framework. Section 3 specifies the novel iBMU method as well as the model search and Markov chain Monte Carlo algorithms used to sample from posterior distributions of interest and approximate posterior summaries. In Section 4, we describe several simulation studies in which the power of the integrative variable selection method is compared with that of a basic BMU method as well as several penalized regression alternatives. Finally, in Section 5, we apply our method to data from the Pharmacogenetics of Nicotine Addiction and Treatment Consortium (PNAT), and Section 6 concludes with a discussion.

2. Bayesian model uncertainty overview

Here, we give an overview of the general BMU framework that is described in more detail in [26, 27]. In particular, we assume that our data is composed of (i) $\mathbf{Y}$, an n-dimensional outcome vector; (ii) $\mathbf{X}$, an $(n \times p)$ dimensional matrix composed of the measured predictors that are included in the model search; and (iii) $\mathbf{Z}$, an $(n \times q)$ dimension matrix composed of the measured confounders that will be forced into every model (such as age and race). Each model $\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}$ is specified by a p-dimensional indicator vector $\boldsymbol{\gamma}$, where $\gamma_j = 1$ if the predictor variable $\mathbf{X}_j$ is included in model $\mathcal{M}_{\boldsymbol{\gamma}}$ and $\gamma_j = 0$ if $\mathbf{X}_j$ is not included in $\mathcal{M}_{\boldsymbol{\gamma}}$. Thus, each model $\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}$ is defined by a unique subset of the p predictor variables of interest.

Given any model $\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}$, we assume that the relation between the outcome variable $\mathbf{Y}$ and the predictor variables can be specified as some generalized linear model with mean $\boldsymbol{\mu}$:

$$\mathcal{M}_{\boldsymbol{\gamma}} : \boldsymbol{\mu} = g^{-1}\left[\beta_0 + \mathbf{Z}\boldsymbol{\beta} + \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right],$$

where g is the link function corresponding to the generalized linear model of interest, β_0 is the intercept common to every model, $\boldsymbol{\beta}$ is the coefficient vector of confounder variables that is also common to every model, $\mathbf{X}_{\boldsymbol{\gamma}}$ is some parametrization of the set of predictors incorporated in model $\mathcal{M}_{\boldsymbol{\gamma}}$, and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the model-specific effect of $\mathbf{X}_{\boldsymbol{\gamma}}$ on the outcome of interest. To simplify notation throughout, we combine all regression coefficients into the vector $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\beta_0, \boldsymbol{\beta}, \boldsymbol{\beta}_{\boldsymbol{\gamma}})$.

2.1. Posterior quantities of interest

The degree to which any model $\mathcal{M}_{\gamma} \in \mathcal{M}$ is supported by the data is quantified by posterior model probabilities defined as

$$p(\mathcal{M}_{\boldsymbol{\gamma}}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathcal{M}_{\boldsymbol{\gamma}})\,p(\mathcal{M}_{\boldsymbol{\gamma}})}{\sum_{\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}} p(Y|\mathcal{M}_{\boldsymbol{\gamma}})\,p(\mathcal{M}_{\boldsymbol{\gamma}})}.$$

The posterior model probabilities are a function of (i) $p(\mathbf{Y}|\mathcal{M}_{\boldsymbol{\gamma}})$, the marginal likelihood of model $\mathcal{M}_{\boldsymbol{\gamma}}$ obtained by integrating out the model-specific parameters $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ with respect to their prior distribution; and (ii) $p(\mathcal{M}_{\gamma})$, the prior probability placed on model $\mathcal{M}_{\gamma}$. Although posterior model probabilities inform us of the models that best explain $\mathbf{Y}$, they do not provide formal marginal inference as to which predictor variables, if any, are associated with $\mathbf{Y}$. To provide formal inference, we calculate marginal posterior inclusion probabilities and marginal Bayes factors (MargBF) for each predictor $\mathbf{X}_j$. The posterior inclusion probabilities are computed as

$$p(\gamma_j = 1|\mathbf{Y}) = \sum_{\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}:\gamma_j = 1} p(\mathcal{M}_{\boldsymbol{\gamma}}|\mathbf{Y}),$$

which is simply the sum of the posterior model probabilities of all models that include $\mathbf{X}_j$ (or all models with $\gamma_j = 1$). The MargBF are defined as the posterior odds divided by the prior odds for inclusion

$$BF\left[\gamma_j = 1 : \gamma_j = 0\right] = \frac{p(\gamma_j = 1|\mathbf{Y})}{p(\gamma_j = 0|\mathbf{Y})} \Big/ \frac{p(\gamma_j = 1)}{p(\gamma_j = 0)}.$$

2.2. Association studies

Although the framework described earlier can be implemented for any generalized linear model, we are interested in the application of case–control association studies. Here, $\mathbf{Y}$ is a binary outcome variable that takes on the value $Y_i = 1$ if individual i is a case and $Y_i = 0$ if individual i is a control. As presented in [19, 28], we use logistic regression to relate $\mathbf{Y}$ to the subset of predictor variables, $\mathbf{X}_{\boldsymbol{\gamma}}$ in model $\mathcal{M}_{\boldsymbol{\gamma}}$:

$$\mathcal{M}_{\boldsymbol{\gamma}} : \text{logit}(\mathbf{Y} = 1) = \beta_0 + \mathbf{Z}\boldsymbol{\beta} + \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}.$$

For the purpose of genetic association studies involving common variants, $\mathbf{X}_{\boldsymbol{\gamma}}$ can be defined to specify the genetic parametrization of each variant included in model $\mathcal{M}_{\boldsymbol{\gamma}}$ as in [19]. For the analysis of rare variants, $\mathbf{X}_{\boldsymbol{\gamma}}$ can define a risk index of the rare variants included in model $\mathcal{M}_{\boldsymbol{\gamma}}$ as in [28].

Given a prior specification for $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$, we must obtain the marginal likelihood to calculate posterior qualities of interest:

$$p(\mathbf{Y}|\mathcal{M}_{\boldsymbol{\gamma}}) = \int p(\mathbf{Y}|\mathcal{M}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}) p(\boldsymbol{\theta}_{\boldsymbol{\gamma}}) \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{\gamma}}.$$

For logistic regression models, this integral is intractable, and Laplace approximations are commonly used to approximate $p(\mathbf{Y}|\mathcal{M}_{\gamma})$. In the supplementary materials of [19], it is shown that under a normal prior distribution for the model-specific parameters, $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$, the Laplace approximation of the marginal likelihood corresponds to a penalized likelihood of the form

$$p(\mathbf{Y}|\mathcal{M}_{\boldsymbol{\gamma}}) \approx \exp\left[-\frac{1}{2}\left(\mathrm{dev}(\mathcal{M}_{\boldsymbol{\gamma}}) + \mathrm{pen}(\mathcal{M}_{\boldsymbol{\gamma}})\right)\right],$$

where $\mathrm{dev}(\mathcal{M}_{\boldsymbol{\gamma}})$ is the deviance of model $\mathcal{M}_{\boldsymbol{\gamma}}$ and $\mathrm{pen}(\mathcal{M}_{\boldsymbol{\gamma}})$ is a penalty on model size that is induced by the choice of variance in the normal distribution. In particular, we will approximate the marginal likelihood with the Akaike information criterion, which roughly corresponds to a prior standard deviation of any standardized log odds ratio (OR) of approximately 2.5.

3. Integrative model uncertainty

We wish to extend the basic BMU framework with iBMU that allows external information to guide the selection of predictors. In particular, we incorporate external information in the estimation of marginal inclusion probabilities and in turn model uncertainty probabilities by introducing a second-stage regression on the probability that any predictor variable $\mathbf{X}_j$ is associated. This model incorporates a set of c predictor-level covariates that is specified in an $(p \times c)$ dimensional matrix $\mathbf{W}$ and that quantify external information on the relationships between the p predictors. In particular, we use a probit model to relate the c predictor-level covariates for predictor $\mathbf{X}_j$ within the vector $\mathbf{W}_j$ to the probability that each predictor is associated by introducing a latent vector $\mathbf{t}$. Each element of $\mathbf{t}$ is distributed normally as

$$t_j|\boldsymbol{\alpha} \sim N(\alpha_0 + \mathbf{W}_j'\boldsymbol{\alpha}, 1).$$

The inclusion indicator of the predictor variable $\mathbf{X}_j$ in model $\mathcal{M}_{\boldsymbol{\gamma}}$ is then specified by the function $\gamma_j = I[t_j > 0]$. Here, $\boldsymbol{\alpha}$ is a c-dimensional regression coefficient that quantifies the increase or decrease in probability that each variable, $\mathbf{X}_j$, is associated based on the c predictor-level covariates, and α_0 specifies the baseline probability of association common to all of the predictor variables. We define α_0 on the basis of the multiplicity corrected model space priors introduced in [19] such that the probability of the null hypothesis, H_0, that no predictors are associated is equal to the probability of the alternative hypothesis, H_A, that at least one predictor is associated with the outcome of interest at baseline (when $\boldsymbol{\alpha} = 0$ for all c covariates). This leads to setting $\alpha_0 = \Phi^{-1}\left(2^{-1/p}\right)$. Finally, to complete the specification of the second-stage model, we assume that $\boldsymbol{\alpha}$ has a prior distribution of $\boldsymbol{\alpha} \sim N(0, \mathbf{I}_c)$. We note that when $\boldsymbol{\alpha} = 0$ for all c covariates, iBMU is equivalent to BMU.

On the basis of our specification, the marginal inclusion probabilities, π_j, can be written as

$$\begin{aligned}\pi_j|\boldsymbol{\alpha} &= \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(t_j-\mu_j)}{2}} \mathrm{d}t_j; \\ &= 1 - \Phi\left(\alpha_0 + \mathbf{W}_j'\boldsymbol{\alpha}\right).\end{aligned}$$

Thus, the probability for each model in the model space $\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}$ given $\boldsymbol{\alpha}$ is

$$p(\mathcal{M}_{\boldsymbol{\gamma}}|\boldsymbol{\alpha}) = \prod_{j=1}^{p} \left[1 - \Phi(\alpha_0 + \mathbf{W}_j'\boldsymbol{\alpha})\right]^{\gamma_j} \left[\Phi(\alpha_0 + \mathbf{W}_j'\boldsymbol{\alpha})\right]^{1-\gamma_j}.$$

3.1. Posterior computation

In many studies, the total number of predictor variables under investigation can be quite large, causing the model space to quickly become innumerable. In these situations, a model search algorithm must be introduced to sample from the space. For our purpose, we use a simple Metropolis–Hastings (MH)

algorithm to sample models from the model space conditional upon $\boldsymbol{\alpha}$. In the MH algorithm, models are evaluated on the basis of the following fitness function:

$$\text{fitness}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|\boldsymbol{\alpha}) = p(\mathbf{Y}|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|\boldsymbol{\alpha}).$$

New models are proposed by randomly selecting one variant and changing the status within the current model. For example, if predictor $\mathbf{X}_j$ is randomly selected and is included in the current model with $\gamma_j = 1$, we propose to remove predictor $\mathbf{X}_j$ from the new model and vice versa. The new model is then accepted with probability $\pi_{\text{accept}} = \min(1, \frac{\text{fitness}(\boldsymbol{\mathcal{M}}_{\text{new}}|\boldsymbol{\alpha})}{\text{fitness}(\boldsymbol{\mathcal{M}}_{\text{old}}|\boldsymbol{\alpha})})$ so that the new model is always accepted if the fitness of it is larger than that of the old model and is accepted with a probability less than 1 if the fitness of the new model is smaller than that of the old.

Given the current sampled model $\boldsymbol{\mathcal{M}}_{\gamma} \in \boldsymbol{\mathcal{M}}$, we use Gibbs sampling techniques to sample from the posterior distribution of $\boldsymbol{\alpha}$ in the second-stage model. The full conditionals that we will need to sample from are calculated as

$$\begin{aligned}
&\boldsymbol{\alpha}|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}, \mathbf{t} \sim N\left(\boldsymbol{\alpha}^*, \mathbf{V}^*_{\boldsymbol{\alpha}}\right); \\
&t_j|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha} \sim N_{(-\infty,0)}(\alpha_0 + \mathbf{W}'_j\boldsymbol{\alpha}, 1) \qquad \text{if } \gamma_j = 0; \\
&t_j|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha} \sim N_{(0,\infty)}(\alpha_0 + \mathbf{W}'_j\boldsymbol{\alpha}, 1) \qquad \text{if } \gamma_j = 1,
\end{aligned}$$

where $\boldsymbol{\alpha}^* = \mathbf{V}^*_{\boldsymbol{\alpha}}[\mathbf{W}'(\mathbf{t} - \alpha_0)]$ and $\mathbf{V}^*_{\boldsymbol{\alpha}} = \left[I_c^{-1} + \mathbf{W}'\mathbf{W}\right]^{-1}$.

By iterating between the MH and Gibbs algorithms, we are able to obtain a sample from the model space, denoted as $\boldsymbol{\mathcal{M}}_s$, and a sample from the posterior distribution of $\boldsymbol{\alpha}$, denoted as $\boldsymbol{\alpha}_s$. Given these samples, we can approximate the posterior probability of each model as follows:

$$p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|\mathbf{Y}) \approx \frac{p(\mathbf{Y}|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|\hat{\pi})}{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}_s} p(\mathbf{Y}|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|\hat{\pi})}, \tag{1}$$

where the posterior model probabilities are renormalized over the sum of sampled models, and $\hat{\pi}$ is the Monte Carlo estimates of the inclusion probabilities given the sampled values $\boldsymbol{\alpha}_s$.

4. Simulation study

To examine the power of iBMU, we have developed a set of 1000 independent simulations composed of 500 cases and 500 controls, and 100 total predictor variables. We assume conditional independence of each of the predictor variables. Also, for each simulation, there is one dichotomous predictor-level covariate that assigns some of the predictors to a single group. For each simulation set, given the simulated predictor variables and predictor-level covariate, we sample a random $\alpha \in \{0, 1, 2, 3\}$ and calculate the probability that each predictor is associated on the basis of the sampled α and the probit equation given in Section 3. We also assume that the baseline α_0 is defined as in Section 3. Specifically, when the total number of predictors is 100, we are assuming that the baseline probability that each predictor is associated is 0.007 (when $\alpha = 0$ and for all predictors with $W_j = 0$). However, when $\alpha \neq 0$, the probability that each variant with $W_j = 1$ is associated increases to 0.072 when $\alpha = 1$, 0.322 when $\alpha = 2$, and 0.705 when $\alpha = 3$. On the basis of these probabilities, we assign anywhere between 0 and 10 (total number of associated predictors within each simulation is randomly assigned) predictors to be associated, and if associated, we assume that they have a modest OR of 1.1.

We also created a set of 1000 genetic association study-based simulations formed by using the genotype data from a systems-based candidate gene study of smoking cessation as part of the PNAT consortium (described in detail in Section 5) [29]. With these simulations, we aim to demonstrate the power of iBMU for more realistic scenarios when correlation exists between the variants within each group (or gene) as well as to show the power and flexibility of the method to account for continuous predictor-level covariates. To create the study-based simulations, $\mathbf{X}$ was formed from genotypes of 122 variants within 789 individuals. The 122 variants are from seven unique gene regions and thus are composed of a great deal of correlation between the markers within each gene. In particular, with the exception of two pairs of variants that are completely correlated, the correlation of the other variants on average across all gene regions ranges from 0.00 to 0.95 with a mean correlation of 0.22. We then create the predictor-level covariate matrix $\mathbf{W}$ by constructing dichotomous dummy variables for the assignment of each variant within a gene region as well as creating a continuous predictor-level covariate. Within

the set of simulations, we select one predictor-level covariate, $\mathbf{W}^*$, from the set of gene dummy variables and continuous covariate to have an increased probability of being associated with the outcome of interest on the basis of randomly assigning an α^* level in $\{0, 1, 2, 3\}$. The corresponding predictor-level covariate and the sampled α^* were then used to assign a probability of association for each marker on the basis of the probit equation given in Section 3. All other predictor-level covariates not equal to $\mathbf{W}^*$ were not used in determining the simulated set of associated markers (or equivalently their α level was assumed to be 0). Also, when α^* was selected to be 0 for the candidate covariate, all associated markers were chosen at random, independent from the predictor-level covariates. Finally, on the basis of these probabilities, we assign anywhere between 0 and 10 predictors to be associated and assumed an OR of 1.5 for all associated markers. The disease status, $\mathbf{Y}$, was then simulated accordingly.

4.1. Comparison with alternative variable selection methods

We compare the power of our novel iBMU approach with the following commonly used variable selection methods:

- Lasso: The lasso [14] introduces an L_1 penalty on the set of regression coefficients to induce sparsity and allow variable selection. Lasso was implemented in *R* using the *glmnet* package [30] that uses coordinate descent to fit the regularization path over a grid of values for the lasso tuning parameter λ.
- Elastic net: The elastic net method [15] is a hybrid of lasso and ridge regression [16] that incorporates both an L_1 and L_2 penalty on the regression coefficients to obtain sparsity and to encourage a grouping effect in that strongly correlated predictors will have similarly estimated regression coefficients. The method was also implemented using the *glmnet* package [30] on a grid of values for λ and the mixing parameter. The optimal mixing parameter was then chosen on the basis of using cross-validation to find the parameter that gives the smallest mean squared error.
- Group bridge: The group bridge [18] allows for variable selection at the group level as well as within groups. The group bridge was implemented with the *grpreg* package in *R* [31] that uses the idea of locally approximated coordinate descent to fit the regularization path over a grid of values for λ. The tuning parameter of the group bridge that induces a penalty on the L_1 norm of the coefficients within a group is selected using cross-validation.
- iBMU and BMU: The novel iBMU approach with one dichotomous predictor-level covariate for the independent simulations and seven dichotomous predictor-level covariates as well as one continuous covariate for the study-based simulations. The effect of the predictor-level covariates, α, is sampled using the Gibbs sampling approach described in Section 3.1. We also implemented the basic BMU framework that is akin to iBMU with $\alpha = 0$ for all predictor-level covariates. Under both methods, we use the MH algorithm described in Section 3.1 to sample models from the model space. Under each method, we run the MH/Gibbs algorithms for 250,000 iterations.

4.2. Variable selection simulation results

For each of the aforementioned methods, we calculate marginal true positive rate (TPR) and false positive rate (FPR) as the proportion of casual and non-causal predictors respectively that are selected using each method. The TPR and FPR are calculated across a grid of thresholds that determine which predictors are selected, and these values are plotted as ROC curves. For the penalized regression methods, we calculate TPR and FPR using varying values of λ as the threshold, and for Bayesian approaches, we calculate the values across varying MargBF thresholds.

Figure 1 plots ROC curves for the independent simulations in which the informativeness of the predictor-level covariate varies from being non-informative to strongly informative on the basis of the truly sampled $\alpha \in \{0, 1, 2, 3\}$. We note that even when we assume that $\alpha = 0$ and the associated predictors are completely independent from the dichotomous predictor-level covariate, there is only a modest reduction in power when using iBMU over BMU, lasso, and elastic net. However, as α increases, there is a substantial increase in the power of iBMU over the other commonly used alternatives that seem to retain the same amount of power across all sets of simulations.

Figure 2 plots ROC curves for the study-based simulations when (i) all predictor-level covariates are assumed to be non-informative, (ii) one gene-based covariate is assumed to be informative, and (iii) the continuous covariate is assumed to be informative. Similarly to the previous simulations, there is an

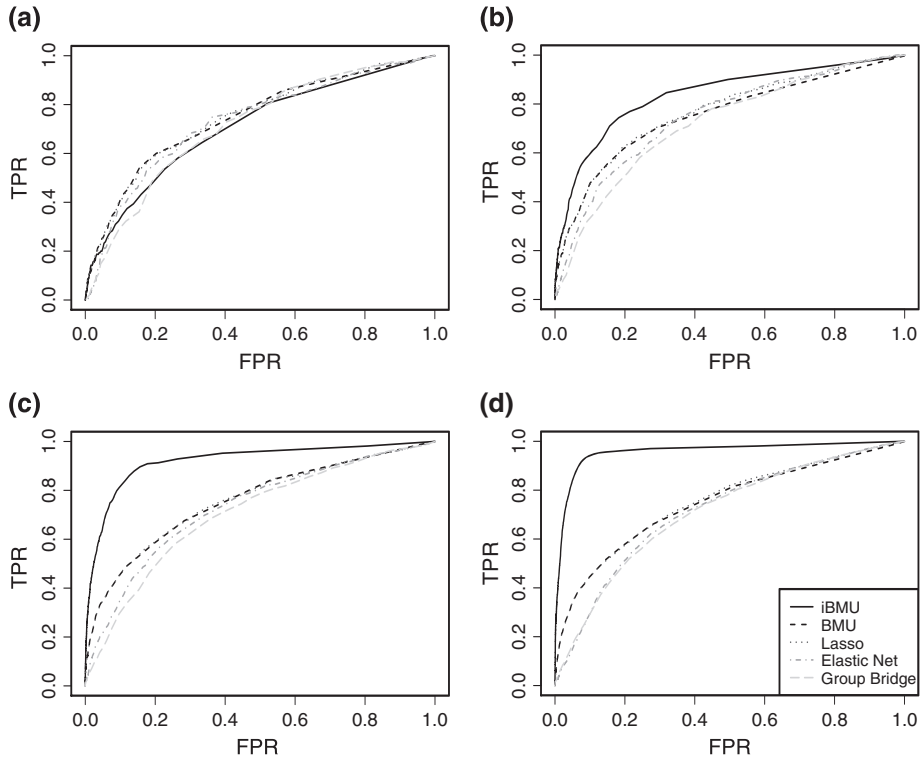

Figure 1. ROC curves are calculated under BMU and iBMU by varying the MargBF threshold that determines which predictor variables are associated with the outcome of interest and are calculated under the penalized regression approaches by varying the value of λ. Given each threshold, the corresponding FPR is plotted on the x-axis and TPR is plotted on the y-axis. Plot (a) corresponds to $\alpha = 0$ in which the predictor-level covariate is non-informative with regard to the associated predictors; and plots (b)–(d) correspond to $\alpha \in \{1, 2, 3\}$, respectively, for varying informativeness of the covariate.

increase in power to detect marginal associations of the variants for iBMU within the informative simulations and little power reduction when the predictor-level covariates are assumed to be non-informative. Also, in the more realistic study-based simulations, we see an increase in power of the BMU approach over all of the penalized regression approaches. Of the penalized regression approaches, elastic net is the most powerful. This is most likely because elastic net takes into consideration the correlation structure of the predictors. Within Table I, we calculate the MargBF threshold needed to achieve FPRs of 0.05 and 0.20 for iBMU and BMU and the λ threshold needed for the penalized regression methods for each simulation. We then report the average MargBF threshold and λ needed as well as the corresponding average TPR under all (i) non-informative simulations, (ii) simulations informed by a dichotomous covariate, and (iii) simulations informed by a continuous covariate. We see that under the iBMU approach, we can achieve an FPR of 0.05 if we use a MargBF threshold of 10 and an FPR of 0.20 if we use a MargBF threshold of 3.

4.3. Sensitivity of marginal Bayes factors

It is of interest to investigate the sensitivity of the estimated MargBF under the iBMU approach that incorporates both continuous and dichotomous predictor-level covariates. With this in mind, Figure 3 plots the log of the MargBF (log(MargBF)) for group informed predictors under iBMU and BMU. Here, group informed predictors are defined as predictors that are assigned to a group based on a dichotomous predictor-level covariate that is assumed to inform the associated predictors (has an $\alpha > 0$). The top plots (a and b) of Figure 3 show the log(MargBF) under iBMU and BMU respectively within the simplistic simulations, and the bottom plots (c and d) show the log(MargBF) within the study-based simulations. Within all plots, the log(MargBF) is plotted as a function of the true OR of each predictor. For each OR, we plot the log(MargBF) for all informed predictors (on the left) and for predictors that are informed by a group where there are only one or two associated members (on the right). Here, we show that

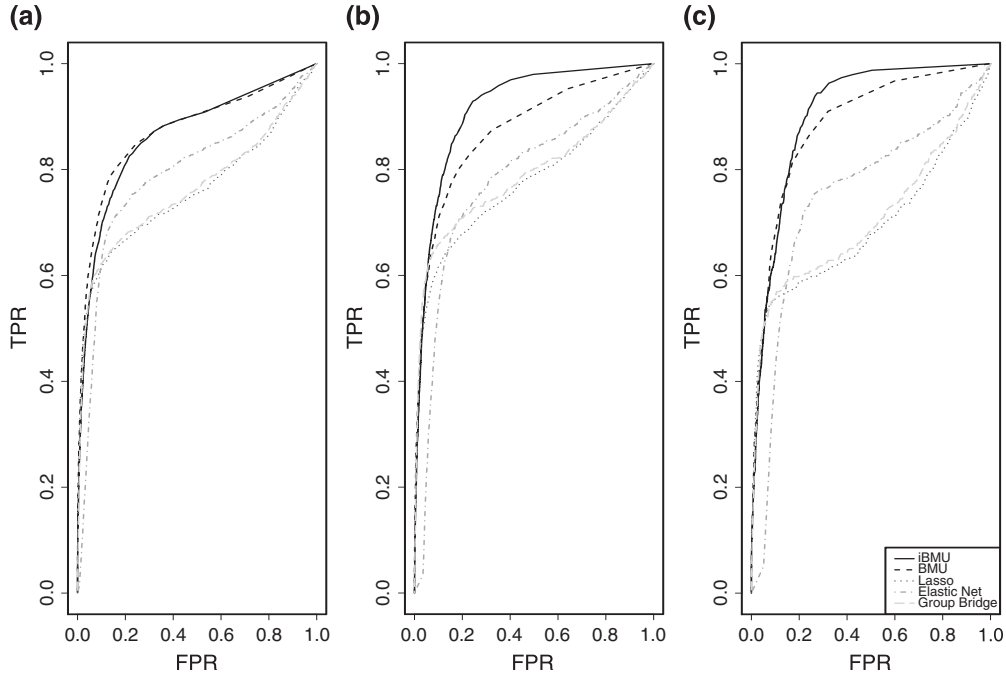

Figure 2. ROC curves are calculated under BMU and iBMU by varying the MargBF threshold that determines which predictor variables are associated with the outcome of interest and are calculated under the penalized regression approaches by varying the value of λ. Given each threshold, the corresponding FPR is plotted on the x-axis and TPR is plotted on the y-axis. Plot (a) corresponds to $\alpha = 0$ for all predictor-level covariates, (b) corresponds to $\alpha > 0$ for an informative gene-based dichotomous covariate, and (c) corresponds to $\alpha > 0$ for the informative continuous covariate.

Table I. Estimated TPR given an FPR of 0.05 and 0.20 for iBMU versus competing methods under (i) all non-informative simulations, (ii) all simulations informed by a dichotomous covariate, and (iii) all simulations informed by a continuous covariate.

	FPR = 0.10				FPR = 0.20			
	Thresh.	$\alpha = 0$	$\alpha_D > 0$	$\alpha_C > 0$	Thresh.	$\alpha = 0$	$\alpha_D > 0$	$\alpha_C > 0$
iBMU	10.00	0.59	0.60	0.51	3.00	0.81	0.88	0.87
BMU	5.00	0.61	0.58	0.50	0.70	0.82	0.81	0.80
Lasso	0.03	0.57	0.52	0.51	0.01	0.62	0.64	0.65
Elastic net	0.20	0.29	0.34	0.38	0.02	0.70	0.67	0.71
Group bridge	0.02	0.52	0.54	0.55	0.01	0.64	0.66	0.68

Also reported are the average MargBF thresholds need to achieve the corresponding FPR for iBMU and BMU, and the average λ thresholds needed for the penalized regression approaches.

although the iBMU approach does lead to an overall increase in log(MargBF) for all predictors within an informed group when compared with the BMU approach, there is a noticeable gap between the average log(MargBF) calculated for associated and non-associated variants within the same informed group (plots (a) and (c)). When we look at the distribution of the log(MargBF) for informed predictors that are within a group with a low number of associated members in plot (a), we do not see an increase in the log(MargBF) of non-associated predictors within the group. However, we do see that the log(MargBF) of the associated variants decreases (although MargBF > 3) such that there are less true positives. This does not appear to be a problem in plot (c) for the PNAT study-based simulations where we assume a larger OR of associated variants.

We are also interested in assessing the sensitivity of the estimated MargBF under stimulations in which a continuous predictor-level covariate informs the associations within our study-based simulations. With this in mind, Figure 4 plots the log(MargBF) under iBMU for all predictors with a continuous covariate

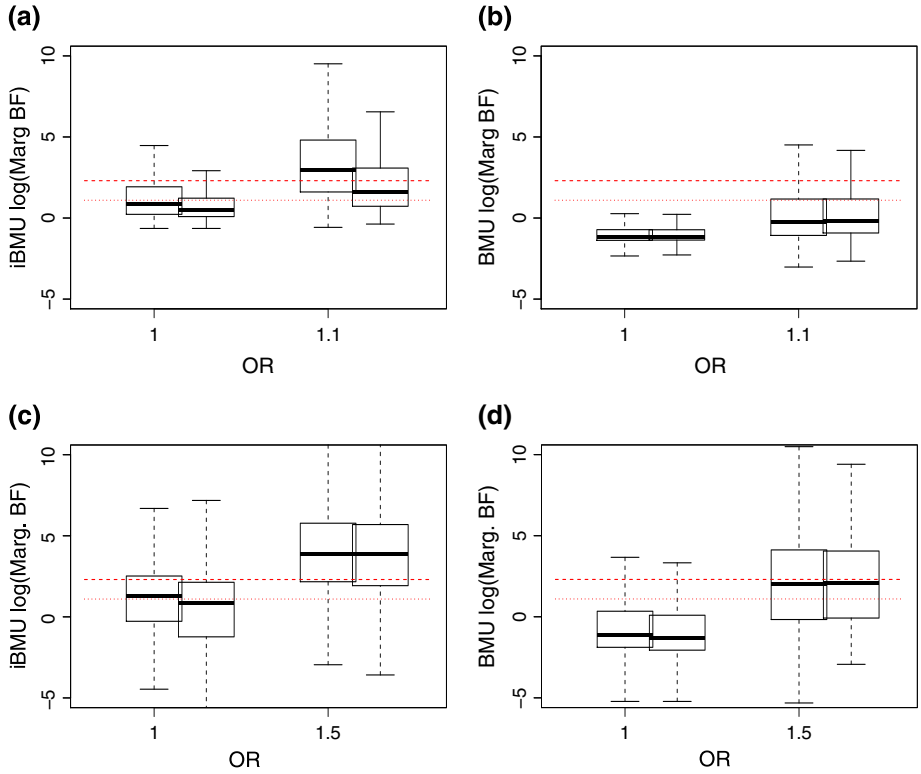

Figure 3. The log(MargBF) for informed predictors are plotted against the true OR for each predictor. Plot (a) corresponds to the log(MargBF) computed under iBMU for the simplistic simulations, (b) under BMU for the simplistic simulations, (c) under iBMU for the study-based simulations, and (d) under BMU for the study-based simulations. For each OR, we plot the log(MargBF) for all informed predictors on the left and for informed predictors that are within a group that has a low number of associated members (one or two). The red lines in each plot correspond to a MargBF threshold of 10 and 3.

less than 1 (the mean of the continuous covariate) in plot (a), greater than 1 but less than 3 (within 2 standard deviations from the mean) in plot (b), and greater than 3 (greater than 2 standard deviations from the mean) in plot (c). Here, we can see that if the continuous covariate is less than 3, a large gap remains in the distribution of the log(MargBF) between associated and non-associated predictors. Furthermore, most of the MargBF of the non-associated predictors remain below the significance thresholds of 3 and 10. However, if the value of the continuous covariate is above 2 standard deviations from the mean, the gap in the distribution of the log(MargBF) between associated and non-associated predictors lessens, and there are a larger number of false positives. Finally, in Figure 4(d), we look at the log(MargBF) under iBMU for all predictors within the group corresponding to gene *CHRNA5*. This gene is of particular interest as there is a high correlation (0.68) between the dichotomous covariate that categorizes the predictors within the gene and the continuous predictor-level covariate. Thus, the value of the continuous covariate for predictors within the gene tends to be high. Here, we see that the gap in the distribution of the log(MargBF) between the associated and non-associated predictors is smaller than that in plot (a) with most of the log(MargBF) for the non-associated predictors below the threshold of 10.

4.4. Model search efficiency simulations results

To explore the model search efficiency of the MH/Gibbs algorithm under iBMU, Figure 5 plots the first iteration of the MH/Gibbs algorithm in which a predictor variable is sampled for both causal and non-causal predictors within the independent simulations. Here, as the simulated α increases and there is a corresponding gain in information via $\mathbf{W}$, the model search becomes more efficient in terms of accepting models with causal predictors earlier on in the stochastic search. Furthermore, as the iteration in which a model with a non-causal variant is first accepted remains constant as the known simulated value of α increases, the increased speed of sampling causal predictors within informative simulations does not come at the cost of also accepting non-causal predictors earlier.

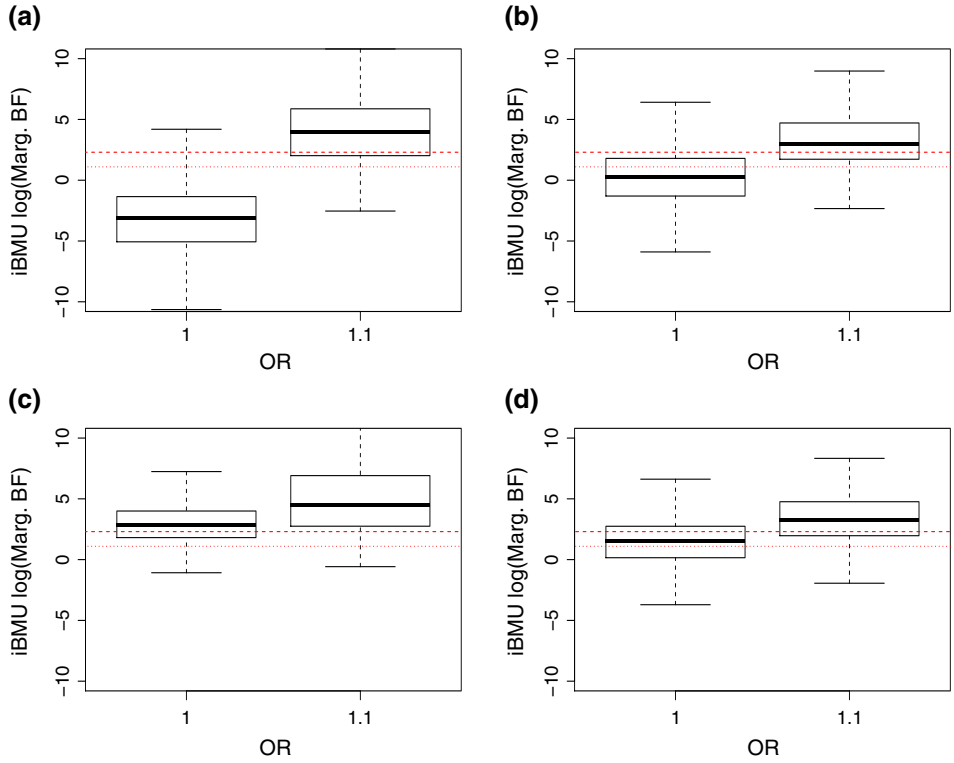

Figure 4. The log(MargBF) is plotted against the true OR for predictors in simulations that assume that associations are informed by a continuous predictor-level covariate. Plot (a) corresponds to the log(MargBF) computed under iBMU for predictors with a continuous covariate less than 1, (b) for predictors with a continuous covariate greater than 1 but less than 3, (c) for predictors with a continuous covariate greater than 3, and (d) for predictors within the group corresponding to gene *CHRNA5*.

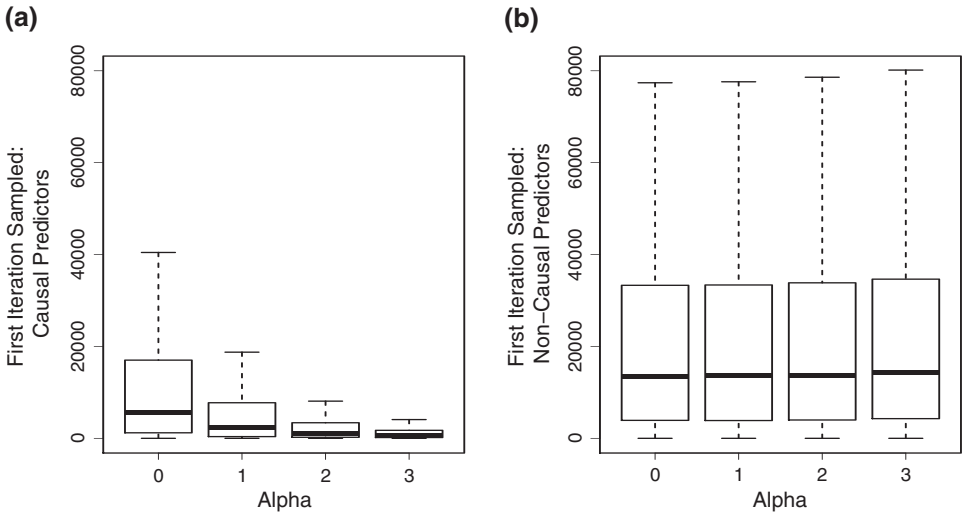

Figure 5. Number of iterations under iBMU until the first acceptance of the causal and non-causal predictors as a function of the simulated α for independent simulations.

5. Genetic association study of smoking cessation

To demonstrate the applicability of iBMU, we analyzed data from a systems-based candidate gene study of smoking cessation as part of the PNAT consortium [29, 32]. The study combines data from two comparable pharmacogenetic trials of smoking cessation treatment conducted by the University of

Pennsylvania Transdisciplinary Tobacco Use Research Center. One aim of the study is to investigate the influence on smoking cessation (abstinence rates at the end of treatment and after a 6-month follow-up period) of variants within genes in the neuronal nicotinic receptor and dopamine systems, studied within a bupropion placebo-controlled randomized clinical trial and a randomized clinical trial comparing transdermal nicotine replacement therapy (patch) to nicotine nasal spray (spray). Detailed descriptions of the study design for the clinical trials were previously reported in [29, 32–34]. Our analysis was limited to 789 persons with self-reported European ancestry. For illustrative purposes to highlight the specific influence of prior covariates, we focused on investigating possible associations within several gene regions related to nicotine processing in the body. This includes 121 single-nucleotide polymorphisms (SNPs) within six unique gene regions that code for several nicotinic acetylcholine receptors. These receptors are involved in the dopamine reward system, a system that is stimulated by nicotine. In addition, we examine genetic variants found within the gene *CYP2A6* that have been previously found to be associated with altered nicotine metabolism. *CYP2A6* converts 80–90% of nicotine to cotinine and subsequently metabolizes cotinine to 3-hydroxycotinine. Additionally, we have measured the nicotine metabolite ratio (NMR), a ratio of 3-hydroxycotinine to cotinine, on all individuals. NMR has been shown to be a stable phenotypic measure of nicotine metabolism, and related associations for genetic variants can serve as a biologically informative prior covariate especially for *CYP2A6* as it is directly involved in nicotine metabolism. As described in [35], we collapse all variants genotyped within *CYP2A6* to a single covariate. The outcome of interest is abstinence after a 6-month follow-up period post-treatment. Finally, our analyses were adjusted for treatment, age, gender, and individual scores from the Fagerström Test for Nicotine Dependence (FTND) by forcing these covariates into all models.

Of particular interest in our analysis is the incorporation of additional biological covariates to aid in the selection of important genetic factors associated with smoking cessation. One such covariate is the categorization of SNPs within gene regions based on our expectations that highly correlated variants within the same gene will show similar evidence of association. Thus, while the model selection aims at identifying single SNPs driving association, structuring the prior covariates to reflect gene regions will allow SNPs within a region to influence the probability of inclusion for other SNPs within that region

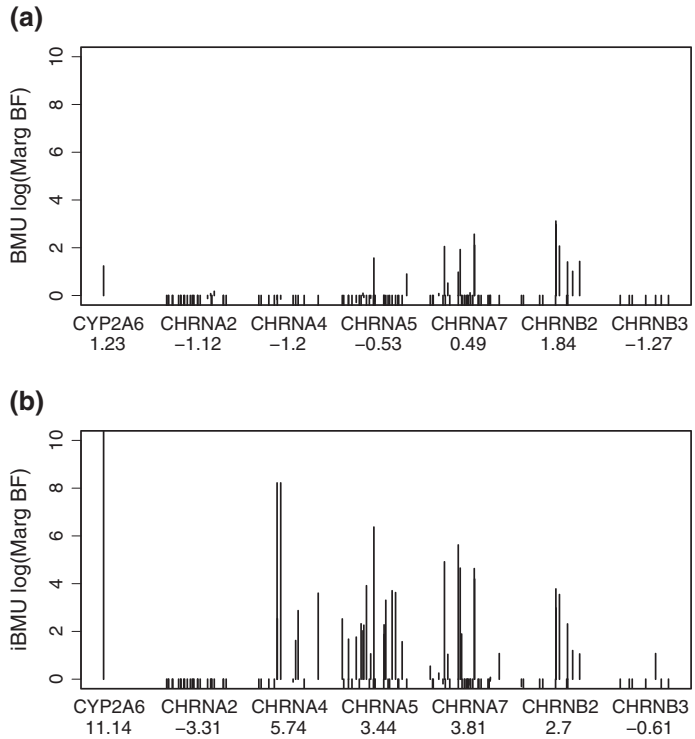

Figure 6. Each plot reports the log(MargBF) of each variant on the y-axis. The variants are organized by gene on the x-axis, and the space between each variant on the x-axis within each gene is proportional to the chromosomal position. Plot (a) corresponds to log(MargBF) calculated under BMU method and plot (b) under the gene region and NMR biomarker-based iBMU method. Gene log(BFs) are reported on the x-axis under each gene.

and provide a summary of the overall influence of the gene via the estimated α. In addition, results from the use of these covariates will also reflect how any single associated SNP can influence the total number of SNPs (ranging from 13 to 33). To reflect potentially more biologically relevant prior information, we construct an additional continuous covariate that is a function of empirical associations of each variant to the NMR. As it is very likely that variants associated with NMR will be more likely to influence smoking cessation, we allow for the degree to which a variant is associated with NMR to inform the prior probability that the variant will be associated with smoking cessation. Specifically, for each variant under consideration (including the covariate coded for *CYP2A6*), we calculate the marginal t-statistic quantifying the likelihood that each variant is associated with NMR. We then use these t-statistics as an additional prior covariate.

To determine the impact of incorporating informative predictor-level covariates to the analysis, we applied BMU and iBMU with gene grouping and NMR–SNP associations as predictor-level covariates to the PNAT data. MargBF were calculated under both methods based on 500,000 iterations of the combine MH/Gibbs techniques described in Section 3 (for BMU only, the MH model search technique was needed). Convergence was determined on the basis of investigating MargBF from two independent runs of the MH/Gibbs algorithms for each scenario. Gene BFs, defined as the posterior odds that at least one variant from gene G_k was associated with smoking cessation divided by the prior odds, were also calculated under each prior scenario. Figure 6 plots MargBF for all variants in the study. Each variant is grouped by gene on the x-axis, and the position of the variant within the gene is relative to the chromosomal position of the variant. Plot (a) is under BMU and plot (b) under iBMU. Gene log(BFs) are reported under each gene on the x-axis. In Table II, we provide results for the top 20 variants within the top five genes ranked under the iBMU approach. Under each gene, we report the estimated effect (α) for the predictor-level covariate corresponding to each gene. We also note that the estimated effect for the NMR-based predictor-level covariate is 0.79. For each variant, we report the variant specific t-statistics for NMR–SNP associations, log(MargBF) under iBMU and BMU, and marginal prior probabilities under iBMU and BMU. Also, for the alternative penalized regression methods, we performed

Table II. Top 20 variants within top five genes.

			iBMU		BMU		Penalized approach		
Gene	SNP	NMR	log (MargBF)	Prior	log (MargBF)	Prior	Group	Elastic net	Lasso
CYP2A6 ($\hat{\alpha} = 0.10$)	CYP2A6	6.39	11.21	0.84	1.23	0.01	1	1	1
CHRNA4 ($\hat{\alpha} = 0.56$)	rs1044396	1.34	8.43	0.24	−0.15	0.01	1	1	0
	rs3787137	1.33	8.43	0.24	−0.16	0.01	0	0	0
	rs4809549	2.11	3.65	0.22	−1.03	0.01	0	0	0
	rs2273505	1.74	2.89	0.32	−1.02	0.01	1	1	0
	rs3787138	1.95	2.58	0.37	−1.53	0.01	0	0	0
CHRNA7 ($\hat{\alpha} = 0.75$)	rs6494211	1.05	5.64	0.20	0.97	0.01	1	1	1
	rs4779969	1.15	4.91	0.22	2.05	0.01	1	1	1
	rs8033518	0.96	4.65	0.18	1.92	0.01	1	1	0
	rs16956223	0.57	4.61	0.12	2.56	0.01	1	1	1
	rs1392808	0.59	4.21	0.12	2.01	0.01	0	0	0
CHRNA5 ($\hat{\alpha} = -0.61$)	rs3743077	2.91	6.39	0.27	1.56	0.01	1	1	1
	rs514743	2.73	3.94	0.23	−1.05	0.01	0	0	0
	rs7178270	2.97	3.72	0.28	−1.64	0.01	0	0	0
	rs950776	2.69	3.62	0.41	−0.90	0.01	1	1	0
	rs1878399	3.02	3.31	0.29	−0.67	0.01	0	0	0
	rs4275821	3.09	2.52	0.31	−0.99	0.01	0	0	0
CHRNB2 ($\hat{\alpha} = 0.27$)	rs2072660	0.67	3.74	0.07	3.11	0.01	1	1	1
	rs3811450	1.09	3.56	0.12	2.07	0.01	1	1	1
	rs2072661	0.83	2.97	0.09	2.96	0.01	0	1	0

Genes are ranked on the basis of the gene BF, and variants within each gene are ranked on the basis of their MargBF under iBMU. Under each gene, we report the estimated effect (α) for the gene-based predictor-level covariate. For each variant, we report (i) rs number, (ii) NMR by variant marginal t-statistic, (iii) iBMU log(MargBF), (iv) iBMU Marg Prior, (vi) BMU log(MargBF), and (vi) BMU Marg Prior; and if the variant was determined to be associated under (vii) group bridge, (viii) elastic net, and (ix) lasso.

k-fold cross-validation to determine the optimal value of λ with regard to mean squared error and report the variants that were determined to be associated under each approach.

By allowing the gene structure and NMR–SNP associations of the variants to inform the prior probability of a marginal association, the approach is able to detect various variants within the *CHRNA4*, *CHRNA5*, *CHRNA7*, and *CHRNB2* gene regions that are likely associated with smoking cessation as well as *CYP2A6*. Without accounting for these biological predictor-level covariates, evidence of associations of variants within these regions is sparse and modest at best. One particular gene of interest is *CHRNA4*. Variants within this gene have been shown in independent studies to be associated with nicotine dependence [36]. Under the BMU and lasso penalized regression approaches, we do not determine that any of the variants within this gene are association. However, when we incorporate gene and correlation structure within the elastic net and group bridge penalized regression approaches, we deem two variants within this region to be associated. Additionally, when we incorporate NMR as well as gene structure within the iBMU approach, there is strong evidence that five variants within this region are associated. Another gene of interest is *CHRNA5*. Although we do not estimate an increase in prior probability of a variant being associated solely on the basis of it being in the gene ($\hat{\alpha} = -0.61$), we do see several of the variants within this gene among the top 20 variants in Table II. This is most likely due to their high values of NMR by variant t-statistic. Variants within this gene are a good example of the ability of iBMU to discern likely associated variants from likely non-associated variants within a gene that is highly correlated with the continuous predictor-level covariate of NMR by variant t-statistic. This is also a good example of the difference between the empirically estimated effects of the gene-based predictor-level covariates on the prior probability of association within the iBMU approach and the posterior gene BFs that give the weight of evidence that at least one variant within a gene is associated.

6. Discussion

The BMU framework provides an extremely powerful and flexible basis for variable selection problems. We have shown that the incorporation of informative predictor-level covariates within this framework leads to an increase in power to detect marginal associations and a more efficient model search algorithm, even when the informativeness is moderate over more commonly used variable selection techniques. By incorporating biological covariates on the gene structure and SNP–NMR associations within the PNAT study, we show strong evidence of an association with variants in *CHRNA4*, *CHRNA5*, *CHRNA7*, *CHRNB2*, and *CYP2A6* and in smoking cessation. Without the incorporation of these prior covariates, the posterior evidence of a marginal association for a variant within any of the gene regions of interest is modest at best. The PNAT analysis was adjusted for treatment, age, gender, and individual scores from the FTND by forcing these covariates into all models. Once we adjusted for the possible confounding variable, we focused on identifying main effects within the variants of interest. It is of future interest to also investigate possible gene-treatment interactions (i.e., placebo, bupropion, patch, and spray).

The integrative variable selection approach described herein has vast implications not only in genetic association studies but also in a wide range of model choice and variable selection problems in a diverse group of interdisciplinary fields. The current implementation focuses on model uncertainty within a logistic regression framework. However, iBMU can easily be extended to other regression problems such as those with a continuous or survival outcome. Within the logistic regression framework, we assume a basic normal prior on the model-specific parameters. However, other prior distributions on the model-specific coefficients can be incorporated into the framework. In particular, predictor-level covariates can also be included in the prior on the coefficient of each included predictor to inform the estimation of the magnitude of the effect of each of the associated predictors. With this in mind, it is of interest in future work to explore the implications of incorporating informative predictor-level covariates on both estimation and inference via model selection.

The BMU and iBMU approaches come at a computational cost of running MH and Gibbs algorithms to sample from the high-dimensional model space and to sample the effects of the predictor-level covariates. The computational complexity of the MH algorithm needed for high-dimensional applications of both BMU and iBMU is a function of the computational cost of estimating model-specific parameters and marginal likelihoods for each unique model sampled, which scales linearly with n and cubically with model size. Thus, an increase in sample size will not cause a significant increase in computation time of the MH algorithm. However, as the model size of sampled models increases, the computation time of the algorithm will increase substantially. The computational complexity of the Gibbs algorithm needed

under the iBMU approach to sample the effects (α) of the predictor-level covariates scales linearly with respect to the number of predictor-level covariates (c) and the total number of predictor variables of interest (p). Therefore, as these parameters increase, we will not see a significant increase in computation time per iteration of the Gibbs sampling algorithm. As an example of the computational cost of the BMU and iBMU algorithms, performing 100,000 iterations of the current MH/Gibbs algorithm under the iBMU approach on the PNAT data took approximately 2 h to complete on a single processor. This can be compared with taking approximately 1.5 h to perform 100,000 iterations of the MH algorithm under the BMU approach on the same processor. The added computational cost of iBMU over BMU is due in part to the added computational complexity of the Gibbs algorithm to sample the effects of the predictor-level covariates. However, it is more likely due to larger models being sampled when the marginal prior probabilities increase for informed predictor variables under the iBMU approach. The computational complexity of BMU and iBMU can be compared with that of the alternative penalized regression approaches, which took approximately 3 s, 6 min, and 30 min to run lasso, elastic net, and group bridge approaches, respectively.

These examples demonstrate the computational complexity of BMU and iBMU approaches for a set number of iterations of the MH and Gibbs algorithms. However, as the total number of predictor variables of interest increases, the number of iterations of the algorithms needed for convergence of posterior quantities will also need to increase. To determine the number of iterations needed for convergence, we suggest doing two independent runs of the algorithms and comparing the global and marginal posterior quantities computed under a set number of iterations of each independent run to determine if the algorithm has converged. The current framework uses a simple MH algorithm to sample models of interest from the innumerable model space. The proposal distribution within the algorithm selects a single predictor variable at random and proposes to mutate the status of the variable in the current model, that is, a random walk. Within this framework, the information within the predictor-level covariates comes into play only in the acceptance probability of the proposed model. However, we have shown that even when the information within the predictor-level covariates is modest, the random walk MH/Gibbs algorithm is more efficient in selecting casual variants over non-causal variants than when there is no prior information. Therefore, the number of iterations needed to reach convergence of the algorithm may be less than that of the MH algorithm under the basic BMU approach. It is of future interest to explore alternative model search algorithms that also incorporate these predictor-level covariates in the proposal distribution to increase the efficiency of the model search even further. Finally, we have investigated the estimation of marginal inclusion probabilities using both a Monte Carlo approach (calculating the proportion of times a variable is sampled) as well as our current approach of calculating them using renormalized posterior model probabilities. We have found that marginal inclusion probabilities calculated from renormalized posterior model probabilities were equally as powerful as those calculated from Monte Carlo estimates.

For ease of specification of informativeness in terms of sensitivity/specificity and for interpretation of the corresponding effect estimates, we have focused the simulations on incorporating dichotomous predictor-level covariates that specify a known group structure of the predictor variables of interest as well as a single continuous predictor-level covariate. However, the amount of information that can be incorporated into an analysis via covariates is extremely flexible. This is demonstrated in our analysis of the PNAT study where the degree to which a variant is associated with NMR (a biomarker that quantifies the rate of nicotine metabolism within an individual) was included as a predictor-level covariate and was shown to significantly inform the prior probability that the variant will be associated with smoking cessation. In particular, for our motivating application of genetic association studies, a vast amount of external biological information exists for the variants under consideration, as discussed in [22]. As one specific example, Cooper and Shendure [37] gave a review of approaches to estimate the overall deleteriousness of genetic variants with the goal of prioritizing disease-causing variants. Many of the reviewed methods use a combination of evolutionary, biochemical, and structural information to guide the estimation. Within our framework, similar information can easily be used as predictor-level covariates; or, once estimated from one of the approaches described in [37], the probability that a variant is deleterious can be used itself as a covariate within the study of interest.

7. Software

Software for the methods described herein is freely available for R within the BVS package on CRAN at the following link: http://cran.r-project.org/web/packages/BVS/.

Acknowledgements

This work has been partially supported by The National Institute of Health (grants R01 ES016813 and R01 ES019876 from NIEHS; U01-DA020830 from NIDA,NCI, NIGMS, and NHGRI; and R21HL115606 from NHLBI and R01CA140561). The authors would like to thank members of the PNAT consortium for use of their data and Duncan Thomas and Paul Marjoram for their helpful critiques.

References

1. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, *et al.* Gene ontology: tool for unification of biology. The gene ontology consortium. *Nature Genetics* 2000; **25**(1):25–29.
2. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Research* 2008; **36**:D480–4. DOI: 10.1093/nar/gkm882.
3. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* 2012; **40**:D742–D753. DOI: 10.1093/nar/gkr1014.
4. Haw RA, Croft D, Yung CK, Ndegwa N, D'Eustachio P, Hermjakob H, Stein LD. The Reactome BioMart database. *Journal of biological databases and curation* 2011. DOI: 10.1093/database/bar031.
5. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the gene ontology consortium. *Nucleic Acids Research* 2010; **38**:D204–D210. DOI: 10.1093/nar/gkp1019.
6. Capanu M, Orlow I, Berwick M, Hummer AJ, Thomas DC, Begg CB. The use of hierarchical models for estimating relative risks of individual genetic variants: an application to a study of melanoma. *Statistics in Medicine* 2008; **27**:1973–1992.
7. Conti DV, Witte JS. Hierarchical modeling of linkage disequilibrium: genetic structure of spatial relations. *American Journal of Human Genetics* 2003; **72**(2):351–363.
8. Greenland S. Hierarchical regression for epidemiologic analyses of multiple exposures. *Environmental Health Perspectives* 1994; **102**(Suppl 8):33–39.
9. Greenland S. Multilevel modeling and model averaging. *Scandinavian Journal of Work, Environment and Health* 1999; **25**(Suppl 4):43–48.
10. Greenland S. Principles of multilevel modelling. *International Journal of Epidemiology* 2000; **29**(1):158–167.
11. Heorn EA, O'Dushlaine C, Segurado R, Gallagher L, Gill M. Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data. *Biostatistics* 2011; **12**(3):445–461.
12. Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P, Witte JS. Using hierarchical modeling in genetic association studies with multiple markers: application to a case–control study of bladder cancer. *Cancer Epidemiology Biomarkers and Prevention* 2004; **13**(6):1013–1021.
13. Thomas DC, Witte JS, Greenland S. Dissecting effects of complex mixtures: who's afraid of informative priors. *Epidemiology (Cambridge Mass)* 2007; **18**(2):186–190.
14. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Statistical Society: Series B* 1996; **58**:267–288.
15. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Roy Statistical Society: Series B* 2005; **67**(2):301–320.
16. Hoerl A, Kennard R. Ridge regression. *Encyclopedia of Statistical Sciences* 1988; **8**:129–136.
17. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J Roy Statistical Society: Series B* 2006; **68**(1):49–67.
18. Huang J, Ma S, Xie H, Zhang C. A group bridge approach for variable selection. *Biometrika* 2009; **96**(2):339–355.
19. Wilson MA, Iversen ES, Clyde MA, Schmidler SC, Schildkraut JM. Bayesian model search and multilevel inference for SNP association studies. *Annals of Applied Statistics* 2010; **4**(3):1342–1364.
20. Chipman H. Bayesian variable selection with related predictors. *Canadian Journal of Statistics* 1996; **24**(1):17–36.
21. Conti DV, Cortessis V, Molitor J, Thomas DC. Bayesian modeling of complex metabolic pathways. *Human Heredity* 2003; **56**(1-3):83–93.
22. Conti DV, Lewinger JP, Tyndale RR, Benowitz NL, Swan GE, Thomas PD. Using ontologies in hierarchical modeling of genes and exposure in biological pathways. *NCI Monographs* 2009; **20**:539–584.
23. Stingo FC, Chen YA, Tadesse MG, Vannucci M. Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. *Annals of Applied Statistics* 2011; **5**(3):1978–2002.
24. Baurley JW, Conti DV, Gauderman WJ, Thomas DC. Discovery of complex pathways from observational data. *Statistics in Medicine* 2010; **29**:1998–2011.
25. Li F, Zhang NR. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of American Statistical Association* 2010; **105**(491):1202–1214.
26. Clyde M, George EI. Model uncertainty. *Statistical Science* 2004; **19**:81–94.
27. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial (with discussion). *Statistical Science* 1999; **14**(4):382–401. Corrected version at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.
28. Quintana MA, Bernstein JL, Thomas DC, Conti DV. Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. *Genetic Epidemiology* 2011; **35**:638–649.
29. Conti DV, Lee W, Li D, Liu J, Berg DVD, Thomas PD, Bergen AW, Swan GE, Tyndale RF, Benowitz NL, *et al.* Nicotinic acetylcholine receptor $\beta 2$ subunit gene implicated in a systems-based candidate gene study of smoking cessation. *Human Molecular Genetics* 2008; **17**(18):2834–2848.

30. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 2008; **33**(1):1–22.
31. Breheny P, Huang J. Penalized methods for bi-level variable selection. *Statistics and Its Interface* 2009; **2**:369–380.
32. Lerman C, Jepson C, Wileyto EP, Epstein LH, Rukstalis M, Patterson F, Kaufmann V, Restine S, Hawk L, Niaura R, *et al.* Role of functional genetic variation in the dopamine D2 receptor (DRD2) in response to bupropion and nicotine replacement therapy for tobacco dependence: results of two randomized clinical trials. *Neuropsychopharmacology* 2006; **31**:231–242.
33. Lerman C, Tyndale R, Patterson F, Wileyto EP, Shields PG, Pinto A, Benowitz NL. Nicotine metabolite ratio predicts efficacy of transdermal nicotine for smoking cessation. *Clinical Pharmacology & Therapeutics* 2006; **79**:600–608.
34. Patterson F, Schnoll RA, Wileyto EP, Pinto A, Epstein LH, Shields PG, Hawk LW, Tyndale RF, Benowitz N, Lerman C. Toward personalized therapy for smoking cessation: a randomized placebo-controlled trial of bupropion. *Clinical Pharmacology & Therapeutics* 2008; **84**(3):320–325.
35. Benowitz N, Swan G, Jacob P, Lessov-Schlagger C, Tyndale R. CYP2A6 genotype and the metabolism and disposition kinetics of nicotine. *Clinical Pharmacology & Therapeutics* November 2006; **80**(5):457–467.
36. Hutchison KE, Allen DL, Filbey FM, Jepson C, Lerman C, Benowitz NL, Stitzel J, Bryan A, McGeary J, Haughey HM. CHRNA4 and tobacco dependence: from gene regulation to treatment outcome. *Arch Gen Psychiatry* 2007; **64**(9):1078–1086.
37. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* September 2011; **12**(9):628–640.